A new method for Quantitative Trait Loci Detection

Charles-Elie Rabier, Céline Delmas

A new method for Quantitative Trait Loci detection

Charles-Elie Rabier

Institut de Mathématiques de Toulouse, Toulouse, France.
INRA UR631, Auzeville, France.

Céline Delmas

INRA UR631, Auzeville, France.

Summary. We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL on the interval $[0, T]$ representing a chromosome (a QTL denotes a quantitative trait locus, i.e. a gene with quantitative effect on a trait). We give the asymptotic distribution of this LRT process under the general alternative that there exist m QTL on $[0, T]$. This theoretical result allows us to propose to estimate the number of QTL and their positions using the LASSO. Our method does not require the choice of cofactors contrary to Composite Interval Mapping (CIM). Besides, our method is not affected by interactions.

Keywords: Gaussian process, Likelihood Ratio Test, Mixture models, Nuisance parameters present only under the alternative, QTL detection, χ^2 process.

1. Introduction

We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting Quantitative Trait Loci, so-called QTL (genes influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by $Y_j,\ j = 1, ..., n$, the observations, which we will assume to be independent and identically distributed (iid). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is purely inherited from A while the other (the "recombined" one), consists of parts originated from A and parts originated from B, due to crossing-overs. The Haldane (1919) modelling assumes that crossovers occur as a Poisson process. Using the Haldane (1919) distance and modelling, each chromosome will be represented by a segment $[0, T]$. The distance on $[0, T]$ is called the genetic distance (which is measured in Morgans).
In a famous article, Lander and Botstein (1989) proposed, with the help of genetic markers, to scan the chromosome, performing a likelihood ratio test (LRT) of the absence of a QTL at every location $t \in [0, T]$. It leads to a "likelihood ratio test process" $\Lambda_n(.)$, and then a natural statistic is the supremum of such a process. This method is called "interval mapping". There have been many papers related to the supremum of the LRT process. For example, we can mention Feingold and al. (1993), Churchill and Doerge (1994), Rebaï and al. (1994), Rebaï and al. (1995), Cierco (1998), Piepho (2001), Chang and al. (2009), Rabier (2010).

The problem is that considering the supremum of the process as a test statistic is appropriate when there is only one QTL on the chromosome but it becomes inappropriate when there are several QTL located on the chromosome. Besides, generally geneticists have no intuition if there is one or several QTL segregating on the chromosome. As a consequence,

a more general approach has to be considered. When multiple QTL occur on the same chromosome, they affect simultanously the LRT process. For instance, when two QTL are located in two different marker interval close but not adjacent, a peak is often found between these two marker interval : it is a ghost QTL (Martinez and Curnow (1992)). Jansen (1993) and Zeng (1994) proposed independently the "Composite Interval Mapping", which consists in combining interval mapping on two flanking markers and multiple regression analysis on other markers (Wu and al. (2007)). This way, the QTL not located in the marker interval tested do not affect anymore the LRT process. Their effects are removed due to multiple regression analysis. Howewer, the choice of markers as cofactor is very complicated. It is still an open question today. Until now, there has been no mathematical proof which could help us on how to choose the set of markers rigorously. In this context, the aim of our paper is to propose an alternative to "Composite Interval Mapping", that is to say a new method which does not require the choice of cofactors.

As mentioned before, in Rabier (2010), the authors suppose that there is no more than one QTL on the chromosome (it is located at $t^\star \in [0,T]$). They show that the LRT process is asymptotically the square of a "non linear interpolated process" centered under H_0 (ie. no QTL on the chromosome) and uncentered of a mean function under the alternative. This mean function depends on the QTL effect and its location $t^\star$. In this paper, we generalize these results to the general alternative that there exist m QTL on $[0,T]$ at $t_1^\star, \cdots, t_m^\star$ with additive effects $q_1, \cdots, q_m$.
The main differences between the alternative of only one QTL and the general alternative, is in the distribution of the trait Y. When there is only one QTL at $t^\star \in [0,T]$, the trait Y, conditionally to information brought by genetic markers located on the chromosome, obeys to a mixture model with known weights :

$$p(t^\star) f_{(\mu+q,\sigma)}(.) + \{1 - p(t^\star)\} f_{(\mu-q,\sigma)}(.) \tag{1}$$

where $f_{(\mu,\sigma)}(.)$ denotes a Gaussian density with mean μ and variance σ^2. (μ, q, σ) are the unknown parameters.
When there are m QTL segregating, the distribution of the trait Y, is a mixture of 2^m components of the form :

$$\sum_{\alpha=1}^{2^m} w_\alpha f_{(M_\alpha,\sigma)}(.)$$

where the w_αs and the M_αs are known functions of the unknown parameters μ, m, $t_1^\star$, ..., $t_m^\star$, q_1, ..., q_m.
In this context, we show that under the general alternative, the LRT process is still asymptotically the square of a "non linear interpolated process". Howewer, the mean function depends this time on the number of QTL, their positions and their effects. This theoretical result allows us to propose a new method to estimate the number of QTL and their positions using the LASSO. Note that in this paper, as in Broman and Speed (2002), the focus is mainly on the estimation of the number of QTL and their positions, rather than on the estimation of the QTL effects. Nevertheless, the effects can be obtained easily with the method that we propose.

The originality of our paper is twofold. First, with our asymptotic study of the LRT process under the general alternative, we are now able to explain mathematically some strange situations which happen when we analyze data. Typically, we generally find a ghost QTL

between two true QTL. Secondly, the originality is in the fact that we propose a new method to find QTL. Our method is very easy to implement and does not require the choice of markers as cofactors which is a major drawback of Composite Interval Mapping. Besides, we prove that our method is not affected by interactions. With the help of simulated data, we show that our method performs better than the Composite Interval Mapping which is largely used in the genetic community. We refer to the book of Van der Vaart (1998) for element of asymptotic statistics used in proofs.

2. Model and Notations

The chromosome is the segment $[0,T]$. K genetic markers are located on the chromosome, one at each extremity. $t_1 = 0 < t_2 < ... < t_K = T$ are the locations of the markers. The "genome information" at t will be denoted $X(t)$. The Haldane (1919) model, which assumes that crossovers occur as a Poisson process, can be written mathematically : let $N(t)$ be a standard Poisson process, the law of $X(t)$ is $\frac{1}{2}(\delta_1 + \delta_{-1})$ and $X(t) = (-1)^{N(t)}X(t_1)$. The Haldane (1919) function $r : [0,\ T]^2 \longmapsto \left[0,\ \frac{1}{2}\right]$ is such as :

$$r(t,t') = \mathbb{P}(X(t)X(t') = -1) = \mathbb{P}(|N(t) - N(t')|\ \text{odd}) = \frac{1}{2}\left(1 - e^{-2|t-t'|}\right)$$

$\bar{r}(t,t')$ will be the function equal to $1 - r(t,t')$.
$r(t,t')$ denotes the probability of recombination between two loci (ie. positions) located at t and t'. $\bar{r}(t,t')$ denotes the absence of recombination. Note that a recombination occurs if there is an odd number of crossovers between the two loci.

We are interested in a quantitative trait Y which is affected by several QTL located on the chromosome. m will refer to the number of QTL and q_s to the QTL effect of the sth QTL. Its position will be called $t_s^\star$. We impose $0 < t_1^\star < ... < t_m^\star < T$ and we will suppose that the QTL effects are additives and there is no interaction between them. In this context, the quantitative trait Y verifies :

$$Y = \mu\ +\ \sum_{s=1}^{m} X(t_s^\star)\ q_s\ +\ \sigma\varepsilon$$

where ε is a Gaussian white noise.
Besides, the "genome information" is available only at locations of genetic markers, that is to say at $t_1, t_2, ..., t_K$. We denote by $X_j(t)$ the value of the variable $X(t)$ for the jth observation. So, in fact, our observation on each individual is $(Y_j,\ X_j(t_1),\ ...,\ X_j(t_K))$. These observations are supposed to be iid.

3. LRT process under the alternative of only one QTL located on $[0,T]$ (Rabier (2010))

Before etablishing the general result of this paper, we first should focus on the work of Rabier (2010), that is to say the case where there is only one QTL lying on $[0,T]$ (ie. $m = 1$). It will be a good way to introduce the LRT process and will make the reading of our paper easier. In order to sum up this previous work, we will consider the same elements and notations used by the authors. As said previously, the authors focus on the famous "Interval Mapping" of Lander and Botstein (1989) which consists in scanning the

chromosome, performing a likelihood ratio test (LRT) of the absence of a QTL at every location $t \in [0, T]$.

We consider values of the parameter t that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For $t \in [t_1, t_K] \backslash \mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, ..., t_K\}$, we define t^ℓ and t^r as :

$$t^\ell = sup\{t_k \in \mathbb{T}_k : t_k < t\} \quad , \quad t^r = inf\{t_k \in \mathbb{T}_k : t < t_k\}$$

In other words, t belongs to the "Marker interval" (t^ℓ, t^r). We define $p(t)$ the weight such as $p(t) = \mathbb{P}\left\{X(t) = 1 \middle| X(t^\ell), X(t^r)\right\}$.
By the Bayes rule,

$$\begin{aligned} p(t) = Q_t^{1,1}\ 1_{X(t^\ell)=1} 1_{X(t^r)=1} \ + \ Q_t^{1,-1}\ 1_{X(t^\ell)=1} 1_{X(t^r)=-1} \\ + Q_t^{-1,1}\ 1_{X(t^\ell)=-1} 1_{X(t^r)=1} \ + \ Q_t^{-1,-1}\ 1_{X(t^\ell)=-1} 1_{X(t^r)=-1} \end{aligned} \tag{2}$$

where :

$$\begin{aligned} Q_t^{1,1} = \frac{\bar{r}(t^\ell, t)\ \bar{r}(t, t^r)}{\bar{r}(t^\ell, t^r)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t^\ell, t)\ r(t, t^r)}{r(t^\ell, t^r)} \\ Q_t^{-1,-1} = 1 - Q_t^{1,1} \quad \text{and} \quad Q_t^{-1,1} = 1 - Q_t^{1,-1} \end{aligned}$$

Let $\theta = (q,\ \mu,\ \sigma)$ be the parameter of the model at t fixed and $\theta_0 = (0,\ \mu,\ \sigma)$ the true value of the parameter under H_0. The likelihood of the triplet $\left(Y,\ X(t^\ell),\ X(t^r)\right)$ with respect to the measure $\lambda \otimes N \otimes N$, λ being the Lebesgue measure, N the county measure on $\mathbb{N}$, is $\forall t \in [t^\ell, t^r]$:

$$L(\theta, t) = \left[p(t) f_{(\mu+q,\sigma)}(y) + \{1 - p(t)\}\ f_{(\mu-q,\sigma)}(y)\right]\ g(t) \tag{3}$$

where $g(t)$ is a function independent of θ.
The likelihood $L_n(\theta, t)$ for n observations is obtained by the product of n terms as above. $\hat{\theta} = (\hat{q},\ \hat{\mu},\ \hat{\sigma})$ will be the maximum likelihood estimator (MLE) of θ.
Under H_0, there is no QTL lying on the interval $[0, T]$. Besides, under H_1, it is supposed that there is only one location where the QTL lies (ie. $m = 1$). In order to deal with this alternative, the location of the QTL, $t^\star$ ($t^\star \in [0, T]$), has to be added in the definition of H_1. So, the alternative hypothesis can be written :

$H_{at^\star}$: "the QTL is located at the position $t^\star$ with effect $q = a/\sqrt{n}$ where $a \in \mathbb{R}^\star$ "

In this context, the authors show that the LRT process, $\Lambda_n(.)$, converges weakly to the square of a "non linear interpolated process". It means that the LRT statistics at each point can easily be deduced from the Wald or score statistics calculated at markers positions. Besides, this "non linear interpolated process" is centered under H_0 and uncentered of a mean function $m_{t^\star}(t)$ under $H_{at^\star}$. This mean function depends on the location of the QTL $t^\star$, the position tested t and the parameter a linked to the QTL effect. It is also a "non linear interpolated fonction" (same interpolation as the process). Then, since they suppose that there is only one QTL on $[0, T]$, the authors have a close formula (due to the interpolation) to compute the supremum of $\Lambda_n(.)$.

4. LRT process under the general alternative of m QTL on $[0,T]$

In the previous Section, it has been supposed that there was only one QTL lying on the interval $[0,T]$. As a consequence, the test statistic used was a natural statistic, that is to say the supremum of the process. The interest is now on studying the same process as previously, $\Lambda_n(.)$, but under the presence of several QTL on the interval $[0,T]$. In this case, the goal is not to perform a test anymore, but to be able to run a model selection in order to estimate the number of QTL and their locations.

Let denote $\vec{t}^{\star}$ the quantity refering to the locations of the QTL. $H_{a\vec{t}^{\star}}$ will be the following assumption :

$H_{a\vec{t}^{\star}}$: " there are m QTL located respectively at $t_1^{\star}$, ..., $t_m^{\star}$ and with effect $q_1 = \frac{a_1}{\sqrt{n}}$, ..., $q_m = \frac{a_m}{\sqrt{n}}$ where $(a_1, ..., a_m) \in \mathbb{R}^{m\star}$ "

We remind that we suppose that the QTL effects are additives and that there is no interaction between them. We will consider values t, $t_1^{\star}$, ..., $t_m^{\star}$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions.

4.1. Results

Theorem *With the previous defined notations,*

$$S_n(.) \Rightarrow Z^{\star}(.) \ , \ \Lambda_n(.) \overset{F.d.}{\to} \{Z^{\star}(.)\}^2$$

as n tends to infinity, under H_0 and $H_{a\vec{t}^{\star}}$ where :

- *$S_n(.)$ is the score process for n observations*
- *$\Rightarrow$ is the weak convergence and $\overset{F.d.}{\to}$ is the convergence of finite-dimensional distributions*
- *$Z^{\star}(.)$ is a Gaussian process with unit variance.*
- *$Z^{\star}(.)$ is the continuous and the "non linear interpolated process" such as :*

$$Z^{\star}(t) = \left\{ \alpha(t)\ Z^{\star}(t^{\ell}) \ + \ \beta(t)\ Z^{\star}(t^{r}) \right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

The mean function of $Z^{\star}(.)$:

- *under H_0, $m(t) = 0$*
- *under $H_{a\vec{t}^{\star}}$, $m_{\vec{t}^{\star}}(t) = \left\{\alpha(t)\ m_{\vec{t}^{\star}}(t^{\ell}) + \beta(t)\ m_{\vec{t}^{\star}}(t^{r})\right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$*

The different quantities are :

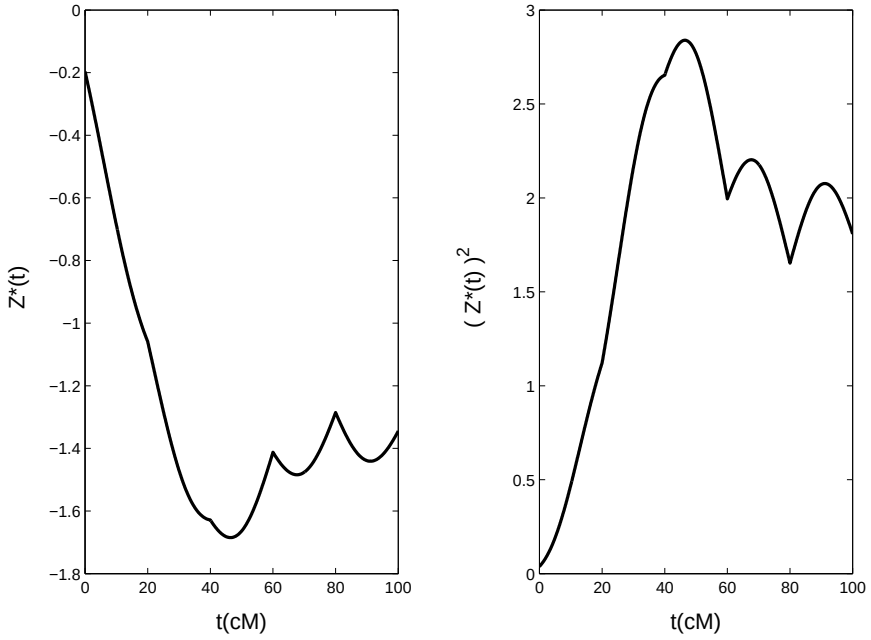
$$\alpha(t) = Q_t^{1,1} + Q_t^{1,-1} - 1, \ \ \beta(t) = Q_t^{1,1} - Q_t^{1,-1}, \ \ Cov\left\{Z(t^{\ell}), Z(t^{r})\right\} = e^{-2(t^r - t^{\ell})}$$

$$m_{\vec{t}^{\star}}(t^{\ell}) = \sum_{s=1}^{m} a_s\ e^{-2\left|t_s^{\star} - t^{\ell}\right|} \ / \ \sigma \ , \ m_{\vec{t}^{\star}}(t^{r}) = \sum_{s=1}^{m} a_s\ e^{-2|t^r - t_s^{\star}|} \ / \ \sigma \ ,$$

$$and \ \ \mathbb{E}\left[\{2p(t)-1\}^2\right] = \{\alpha(t)\}^2 \ + \ \{\beta(t)\}^2 \ + \ 2\ \alpha(t)\ \beta(t) e^{-2(t^r - t^{\ell})}.$$

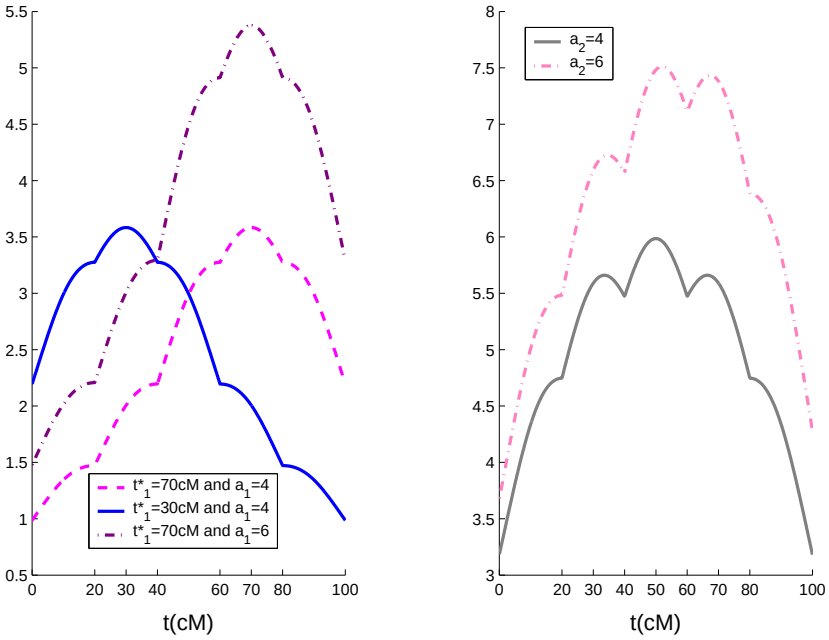
The proof is given in Section 7.1.

4.2. Illustration of the theorem and of the Ghost QTL phenomenon



Process $Z^\star(.)$ Process $\{Z^\star(.)\}^2$

Fig. 1. A path under H_0 of the processes $Z^\star(.)$ and $\{Z^\star(.)\}^2$ ($T = 100$cM, 6 markers equally spaced every 20cM)



$m = 1$ $\qquad$ $m = 2,\ t_1^\star = 30\text{cM},\ t_2^\star = 70\text{cM},\ a_1 = 4$

Fig. 2. Mean function $m_{\vec{t}^\star}(t)$ as a function of the number m of QTL, their positions $t_s^\star$, and the parameters a_s linked to the QTL effects ($T = 100$cM, 6 markers equally spaced every 20cM)

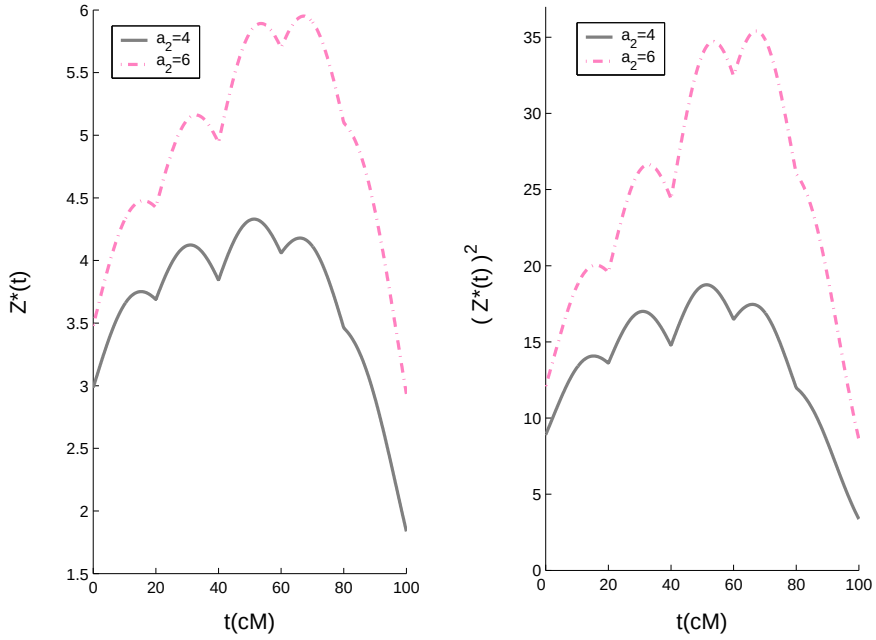


Fig. 3. Same path of $Z^\star(.)$ and $\{Z^\star(.)\}^2$ as under H_0 but under $H_{a\vec{t}^\star}$ ($m = 2$, $t_1^\star = 30$cM, $t_2^\star = 70$cM, $a_1 = 4$, $T = 100$cM, 6 markers equally spaced every 20cM)

In order to illustrate the theorem, we will consider a genetic map which consists of a chromosome of size $T = 100$cM with 6 markers equally spaced every 20cM. Figure 1 refers to the absence of QTL on the chromosome. On the left-side, a path of the process $Z^\star(.)$ is represented under H_0. As there is not any QTL, it corresponds only to noise. Besides, we can observe the interpolation obtained between genetic markers. The same path corresponding to the process $\{Z^\star(.)\}^2$ has been added on the right-side : in genetics, we call this path "a likelihood profile". It is usually this path that we obtain when we analyze data. Note that many authors, instead of computing the process $\Lambda_n(.)$, focus on the LOD process, $LOD_n(.)$,, where $LOD_n(.) = \Lambda_n(.)/\{2\log(10)\}$.

Figure 2 represents the signal. On the left-side, we present some mean functions $m_{\vec{t}^\star}(t)$ when only one QTL ($m = 1$) is located on the chromosome. As expected, the supremum of these interpolated functions is obtained at the location of the QTL. Besides, the larger the QTL effect is, the stronger the signal is. On the right-side, the focus is on $m_{\vec{t}^\star}(t)$ when $m = 2$. According to the theorem, $m_{\vec{t}^\star}(t)$ is obtained by summing the mean functions corresponding to the case $m = 1$. As a consequence, the functions $m_{\vec{t}^\star}(t)$ of the graph of the right-side are easily obtained from those of the graph of the left-side. Let's focus on the curve in solid line. The two QTL are located respectively at $t_1^\star = 30$cM and $t_2^\star = 70$cM. So, the marker interval (40cM, 60cM) is adjacent to the two marker intervals where the QTL are located. As a result, we can observe on the graph that the biggest peak is obtained in the interval (40cM,60cM) and that the supremum is obtained in the middle of this marker interval, at 50cM. Note that it is obtained exactly at 50cM since we consider exactly the same effect ($a_1 = a_2 = 4$) and that there is symmetry due to the location of the QTL and the length of the chromosome. If now we consider a larger effect for the second QTL ($a_2 = 6$) located at $t_2^\star = 70$cM (dashed line), we can observe almost the same two peaks in the intervals (40cM,60cM) and (80cM,100cM). Besides, the supremum of the mean function is obtained at 52cM. It is like a barycenter : some weights are affected to the QTL as a function of their effects, so the signal and the location of the supremum is affected by these weights.

Figure 3 is the analogous of Figure 1 under the alternative of 2 QTL located at $t_1^\star = 30$cM and $t_2^\star = 70$cM. As in Figure 1, the path of the process $Z^\star(.)$ is on the left-side whereas the one corresponding to $\{Z^\star(.)\}^2$ is on the right-side. According to the theorem, in order to obtain the path of $Z^\star(.)$ under $H_{a\vec{t}^\star}$, we have to sum the path of $Z^\star(.)$ under H_0 (ie. the noise), and the mean function $m_{\vec{t}^\star}(t)$ (ie. the signal). In other words, the path of $Z^\star(.)$ under $H_{a\vec{t}^\star}$ has been obtained by adding the path of $Z^\star(.)$ presented in Figure 1 and the mean function of the graph of the right-side of Figure 2. Note that on the right-side of Figure 3, the likelihood profile (ie. the path of $\{Z^\star(.)\}^2$) has easily been obtained by computation of the square of $Z^\star(.)$. We can observe in Figure 3 that, when the effects of the two QTL are the same (ie. the solid lines), the biggest peak is obtained between 40cM and 60cM which is a marker interval where there is no QTL : such a peak is called a ghost QTL (Martinez and Curnow (1992)). It was expected since the supremum of the signal was obtained at 50cM.

Note that when we increase the effect of the second QTL (ie. the dashed lines), the biggest peak is obtained in the marker interval (60cM, 80cM) which is the interval which contains the second QTL. It is due to the noise since the signal is almost the same in the intervals (40cM,60cM) and (60cM,80cM) whereas the values of $Z^\star(.)$ are larger under H_0 in the marker interval (60cM, 80cM) than in the interval (40cM, 60cM).

To conclude, we wanted to highlight here the fact that the likelihood profiles in QTL

detection, are the results of two components : the noise and the signal which contains informations on the number of QTL, their effects and positions. Besides, when two QTL are located in two different markers intervals close but not adjacent, a ghost QTL is often found between these two markers intervals : it is due to the signal (cf. Figure 2). We can only say “often" because of the noise which affects also the likelihood profiles.

5. A new method for QTL detection

In this section, the goal is to propose a method to estimate the number of QTL, their effects and their positions combining results of the theorem and a penalized likelihood method.

5.1. Introducing our method

According to the theorem, if we discretize the score process at markers positions, we have when n is large :

$$\vec{S}_n = \vec{m}_{\vec{t}^\star} \ + \ \vec{\varepsilon}$$

where $\vec{S}_n = (S_n(t_1) \ , \ S_n(t_2) \ , \ ... \ , \ S_n(t_K))' \ , \ \vec{m}_{\vec{t}^\star} = (m_{\vec{t}^\star}(t_1) \ , \ m_{\vec{t}^\star}(t_2) \ , \ ... \ , \ m_{\vec{t}^\star}(t_K))'$ and $\vec{\varepsilon} \sim N(0, \Sigma)$ with $\Sigma_{kk'} = e^{-2|t_k - t_{k'}|}$.
It will be useful to decorrelate the components of $\vec{S}_n$ for running the penalized likelihood method. That's why, we propose to keep only points of the process taken at marker positions : we can perform a Cholesky decomposition of Σ (we remind that S_n is an “interpolated process"). However, we will look for QTL not only on markers postions.
Let consider the Cholesky decomposition $\Sigma = AA'$. It comes :

$$A^{-1}\vec{S}_n = A^{-1}B\left(\frac{a_1}{\sigma} \ , \ ... \ , \ \frac{a_m}{\sigma}\right)' \ + \ A^{-1}\vec{\varepsilon}$$

where B is a matrix of size $K \times m$ such as $B_{ks} = e^{-2|t_k - t_s^\star|}$.
The problem is that the number m of QTL and their positions $t_1^\star$,...,$t_m^\star$ are unknown. So, we consider a new discretization of $[0, T]$ corresponding to all the locations where we think the QTL can be located : $0 \leqslant \tilde{t}_1 < \tilde{t}_2 < ... < \tilde{t}_L \leqslant T$. $\tilde{a}_1, ..., \tilde{a}_L$ will be the corresponding effects divided by σ. As a consequence, we can rewrite the model :

$$A^{-1}\vec{S}_n = A^{-1}\tilde{B}\,(\tilde{a}_1 \ , \ ... \ , \ \tilde{a}_L)' \ + \ A^{-1}\vec{\varepsilon} \tag{4}$$

where $\tilde{B}$ is a matrix of size $K \times L$ such as $\tilde{B}_{kl} = e^{-2|t_k - \tilde{t}_l|}$.
At this time, we would like to know which of the coefficients $\tilde{a}_1, ..., \tilde{a}_L$ are exactly 0 : it will tell us where the QTL are located. As a consequence, a natural approach is to use the LASSO Tibshirani (1996) :

$$argmin_{(\tilde{a}_1,...,\tilde{a}_L)'} \left\| A^{-1}\vec{S}_n - A^{-1}\tilde{B}\,(\tilde{a}_1 \ , \ ... \ , \ \tilde{a}_L)' \right\|^2 \quad \text{provided that } |\tilde{a}_1| + ... + |\tilde{a}_L| \leqslant \zeta$$

ζ is a tuning parameter. It will control the amount of shrinkage that is applied to the estimates Tibshirani (1996). A large (resp. small) ζ will lead to the estimation of a large (resp. small) number of QTL m. We will estimate ζ using cross validation as described in Chapter 7 of Hastie and al. (2001).

5.2. Computing the score and the Wald processes

In order to run our method, we need to calculate the score process discretized at marker locations. We remind that t_k refers to the location of marker k. According to Rabier (2010), the score statistic on marker k verifies :

$$S_n(t_k) = \sum_{j=1}^{n} \frac{(y_j - \mu)\ \left\{2\ 1_{X_j(t_k)=1} - 1\right\}}{\sigma\ \sqrt{n}} \tag{5}$$

According to Prohorov and by contiguity (cf. Section 7.1), the score test can be obtained, replacing μ by $\bar{y} := \sum_{j=1}^{n} y_j/n$ and σ by $\left\{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2\right\}^{1/2}$.
Besides, let $W_n(.)$ the Wald process for n observations. As the model is regular and by contiguity, we have $\forall t \in [0,T]$, $S_n(t) = W_n(t) + o_P(1)$ where $o_P(1)$ is a sequence which converges to 0 in probability under H_0 and $H_{a\vec{t}^\star}$.
As a consequence, our method for QTL detection is also suitable with the Wald process $W_n(.)$ (just replace S_n by W_n in Section 5.1). In this case, according to Rabier (2010) :

$$W_n(t_k) = n\ \hat{q} / \left\{\sum_{j=1}^{n}(y_j - \bar{y})^2\right\}^{1/2}$$

where $\hat{q}$ is the maximum likelihood estimator of q.

5.3. How to improve our method

Our method is based on the asymptotic result of the theorem. As a consequence, we have to consider a number of observations n large enough to run the method. We remind that we have n observations since we consider n individuals. On the other hand, in the model (4), we have this time only K observations which correspond to the score statistic (obtained from the n individuals) on markers and decorrelated. Besides, there are L parameters $\tilde{a}_1$, ..., $\tilde{a}_L$ to estimate (if we except ζ). We remind that $\tilde{t}_1$,... ,$\tilde{t}_L$ denote the location where we are going to look for QTL. In most of cases, as we don't have any idea where the QTL are lying, we will look for QTL on markers and between markers. If we consider d positions in each marker interval, then $L = K(d+1) - d$. It comes $L >> K$. In such a situation, the LASSO is suitable. Howewer, in order to improve the performance of the LASSO, it would be nice if we could deal with a large number of observations K. The problem is that K refers to the number of genetic marker which is constant. So, we have to find an alternative. In an asymptotic study, the question is always the same : how many individuals n are needed to reach the asymptotic ? We have to keep in mind that even if n is very large, we will only deal with K observations (ie. the number of markers) in model (4). As a result, we propose to split the individuals into groups and to analyze these groups separately, that is to say computing the score (or Wald) process for each group. Obviously, we have to deal with a number of individuals large enough in each group in order to reach the asymptotic. We consider groups of same sizes and we call I the number of groups : n/I is the number of individuals in each group. $S_I^i(.)$ denotes the score process for the ith group. According to the theorem, $S_I^i(.)$ is asymptotically the square of a "non linear interpolated process" with a mean function $\vec{m}_{\vec{t}^\star,I}(.)$ under the alternative, verifying

$$m_{\vec{t}^\star,I}(t) = \left\{\alpha(t)\ m_{\vec{t}^\star,I}(t^\ell) + \beta(t)\ m_{\vec{t}^\star,I}(t^r)\right\} / \sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}$$

where

$$m_{\vec{t}^\star,I}(t^\ell) = \sum_{s=1}^{L} a_s \, e^{-2|t_s^\star - t^\ell|} \, / \, (\sigma\sqrt{I}) \quad , \quad m_{\vec{t}^\star,I}(t^r) = \sum_{s=1}^{L} a_s \, e^{-2|t^r - t_s^\star|} \, / \, (\sigma\sqrt{I})$$

Note that $\sqrt{I}$ at the denominator comes from the fact that the QTL effects have been defined as a function of the total number of individuals n.
So, since the groups are independent, we can easily adapt our method of Section 5.1. We have now :

$$\left(\vec{S}_I^1 \, , \, ... \, , \, \vec{S}_I^I\right)' = \left(\vec{m}_{\vec{t}^\star,I} \, , \, ... \, , \, \vec{m}_{\vec{t}^\star,I}\right)' \, + \, (\vec{\varepsilon}_1 \, , \, ... \, , \, \vec{\varepsilon}_I)'$$

where $\vec{m}_{\vec{t}^\star,I} = \left(m_{\vec{t}^\star,I}(t_1) \, , \, m_{\vec{t}^\star,I}(t_2) \, , \, ... \, , \, m_{\vec{t}^\star,I}(t_K)\right)$, $\vec{S}_I^i = \left(S_I^i(t_1) \, , \, S_I^i(t_2) \, , \, ... \, , \, S_I^i(t_K)\right)$ and $\vec{\varepsilon}_i$ iid of size $1 \times K$ such as each $\vec{\varepsilon}_i \sim N(0, \Sigma)$ with $\Sigma_{kk'} = e^{-2|t_k - t_{k'}|}$.
In the same way as previously (cf. Section 5.1) provided that this time $\tilde{a}_1$, ..., $\tilde{a}_L$ are the effects divided by $\sigma\sqrt{I}$:

$$\Gamma \, \left(\vec{S}_I^1 \, , \, ... \, , \, \vec{S}_I^I\right)' = \Xi \, (\tilde{a}_1 \, , \, ... \, , \, \tilde{a}_L)' \, + \, \Gamma \, \vec{\varepsilon} \tag{6}$$

Γ is a square matrix of size KI such as $\Gamma = Diag\left[A^{-1} \, , \, ... \, , \, A^{-1}\right]$.
Ξ is a column vector of components $A^{-1}\tilde{B}$ replicated I times.
To conclude, we propose to use the LASSO Tibshirani (1996) :

$$argmin_{(\tilde{a}_1,...,\tilde{a}_L)'} \left\| \Gamma \, \left(\vec{S}_I^1 \, , \, ... \, , \, \vec{S}_I^I\right)' - \Xi \, (\tilde{a}_1 \, , \, ... \, , \, \tilde{a}_L)' \right\|^2 \quad \text{provided that } |\tilde{a}_1| + ... + |\tilde{a}_L| \leqslant \zeta$$

6. Simulations

In this Section, we perform our method using Wald processes (cf. Section 5.2) and 5 fold cross validation for the LASSO. We consider 100 populations of size $n = 320$. We use mainly MATLAB to perform our method. We used R to perform The LASSO with package LARS of Hastie and Efron. Composite Interval Mapping was performed using (R/qtl Broman and al. (2003)).

6.1. How does our method perform?

In order to illustrate the performances of our method, we consider a sparse map which consists of 6 genetic markers equally spaced every 20cM on a chromosome of length $T =$ 100cM. We look for a QTL every 5cM. In order to make groups, we have to find a good compromise between having enough individuals in each group to reach the asymptotic, and having a large number of groups to increase the performances of the LASSO. We split here our 320 individuals into 8 groups of 40 individuals in order to improve the method (cf. Section 5.3). Indeed, it is reasonable to consider the asymptotic to be reached with 40 individuals (Rabier (2010)). As a consequence, we have now $L = 21$ parameters to estimate with $6 \times 8 = 48$ observations (6 markers and 8 groups).

We study several situations with 2, 3 and 4 QTL. We will say that a QTL is truly identified if the QTL is find in a neighbourhood of 5cM of the true position (ie an interval of length

10cM centered on the true location). Besides, in order to count the number of QTL found, we have choosen not to penalize if several QTL were found in the 10cM intervals centered on the true locations, whereas we have choosen to penalize a lot for any QTL found outside of the intervals. As a consequence, we count only one QTL if 2 or 3 QTL are found in the 10cM intervals centered on the true locations and we count one QTL for every QTL found outside these intervals.

In Figure 4, we study a situation with 2 QTL located on the chromosome. First, two QTL linked in repulsion (ie with opposite signs) are located at positions 10cM and 70cM on the chromosome. We have to keep in mind that as our method is based on contiguity, the QTL effects have to be close to 0. However, we can see in Figure 4, that the method gives good results even when the effects are not so close to 0. Note that the heritability is indicated just for information but it is not linked to the performances of our method since the bigger the effects are the bigger the heritability is. The number of QTL found is slightly greater than 2, but it is reasonable since we penalize a lot when we are outside of the QTL intervals. We obtain the same conclusions for the two QTL linked in coupling (ie. with same signs) presented on the right side of Figure 4. Good performances of the methods are also illustrated in Figure 5 when 3 and 4 QTL are located on the chromosome.

6.2. Comparison with the Composite Interval Mapping

We propose here to compare our method with the Composite Interval Mapping (CIM) of Jansen (1993) and Zeng (1994), largely used in the genetic community. We remind that CIM consists in combining interval mapping on two flanking markers and multiple regression analysis on other selected markers (Wu and al. (2007)). This way, the QTL not located in the marker interval tested don't affect the test statistics anymore. As a consequence, it is possible to perform separately interval mapping in each marker interval to test the presence of a QTL in the interval. However, the choice of the markers as cofactors is very empirical : we don't know how to choice the set of markers in a mathematical point of view.

For the comparison between our method and CIM, we use the same configuration as in Section 6.1. We study several situations with 2, 3 and 4 QTL on the chromosome (see Figures 6 and 7). We compute 4 kinds of CIM. First, we consider two ways of choosing the cofactors : $CIM(20)$ (resp. $CIM(40)$) refers to CIM with markers considered as covariates if they do not belong to a window size of 20cM (resp. 40cM) of the position tested. Secondly, we consider two ways of computing the thresholds : one obtained using 1000 permutations and called $Shuff$ here (Churchill and Doerge (1994)), and another which is obtained theoretically under H_0 (6.76 according to Rabier (2010)).

In order to count the number of QTL for CIM, for each marker interval, we count one QTL if the supremum of the process is above the threshold (it corresponds to the definition of CIM). Besides, for CIM, we will say that a QTL is truly identified if the QTL is find in a neighbourhood of 5cM of the true position. For instance, if a QTL is located at 10cM, the supremum in the marker interval (0cM;20cM) has to be obtained between 5cM and 15cM. Howewer, if we consider a QTL located at 40cM (ie on the third marker), we will consider that this QTL is truly identified if the supremum in the marker interval (20cM;40cM) is obtained between 35cM and 40cM, or if it is obtained between 40cM and 45cM in the marker interval (40cM;60cM).

According to Figure 6, if we consider 2QTL at 10cM and 70cM with effects -0.6 and 0.8, we can see that $CIM_{H_0}(20)$ is the best way to perform CIM : we find 1.84 QTL and the

true QTL are largely found. However, if we consider the same 2 QTL but with effects 0.4 and -0.6, $CIM_{H_0}(20)$ performs badly. $CIM_{Shuff}(20)$ seems to the best way to perform CIM : the true QTL are largely found but we find 3.26 QTL. If we consider 3 QTL, the best way to perform CIM is $CIM_{Shuff}(40)$ but we find 4.97 QTL. As a consequence, the choice of the cofactors and the choice of the thresholds highly depends of the configuration : CIM is very empirical. If now we have a look on our method in Figure 6, we obtain nice results : the QTL are largely found and the number of QTL found is good whatever the configuration studied. Same conclusions hold with 4 QTL (see Figure 7).

6.3. Our method is not affected by epistasis

Until now, we have supposed that the QTL effects were additives and that there were no interaction between them (cf. Section 2). However, there are many interactions between loci in the genome (ie. epistasis). That's why we propose here to integrate interactions in the model considered. We remind that m refer to the number of additive QTL and q_s to the QTL effect of the sth additive QTL. Its position is $t_s^\star$. We will call $\tilde{m}$ the number of interactions and $\tilde{q}_s$ the effect of the sth interaction. The loci corresponding to the sth interaction will be called $\tilde{t}_{2s-1}$ and $\tilde{t}_{2s}$. In this context, the quantitative trait Y verifies :

$$Y = \mu \ + \ \sum_{s=1}^{m} X(t_s^\star)\, q_s \ + \ \sum_{s=1}^{\tilde{m}} X(\tilde{t}_{2s-1}) X(\tilde{t}_{2s})\, \tilde{q}_s \ + \ \sigma\varepsilon$$

where ε is a Gaussian white noise.

We introduce two new hypothesis :

$H_{a\vec{t}^\star,\, b\tilde{t}}$: " there are m additive QTL located respectively at $t_1^\star$, ..., $t_m^\star$ and with effect $q_1 = \frac{a_1}{\sqrt{n}}$, ..., $q_m = \frac{a_m}{\sqrt{n}}$ where $(a_1, ..., a_m) \in \mathbb{R}^{m\star}$
and there are $\tilde{m}$ interactions : between loci $\tilde{t}_1$ and $\tilde{t}_2$, ..., between loci $\tilde{t}_{2\tilde{m}-1}$ and $\tilde{t}_{2\tilde{m}}$, with effects respectively $\tilde{q}_1 = \frac{b_1}{\sqrt{n}}$, ..., $\tilde{q}_{\tilde{m}} = \frac{b_{\tilde{m}}}{\sqrt{n}}$ where $(b_1, ..., b_{\tilde{m}}) \in \mathbb{R}^{\tilde{m}\star}$ ".

$H_{0,\, b\tilde{t}}$: " there is not any additive QTL on $[0, T]$
and there are $\tilde{m}$ interactions : between loci $\tilde{t}_1$ and $\tilde{t}_2$, ..., between loci $\tilde{t}_{2\tilde{m}-1}$ and $\tilde{t}_{2\tilde{m}}$, with effects respectively $\tilde{q}_1 = \frac{b_1}{\sqrt{n}}$, ..., $\tilde{q}_{\tilde{m}} = \frac{b_{\tilde{m}}}{\sqrt{n}}$ where $(b_1, ..., b_{\tilde{m}}) \in \mathbb{R}^{\tilde{m}\star}$ ".

Proposition *Under* $H_{0,\, b\tilde{t}}$ *and under* $H_{a\vec{t}^\star,\, b\tilde{t}}$

$$\forall k \quad S_n(t_k) = Z^\star(t_k) \ + \ o_P(1) \quad \textit{and} \quad \Lambda_n(t_k) = \{Z^\star(t_k)\}^2 \ + \ o_P(1)$$

where $Z^\star(.)$ *is the Gaussian process of the theorem (cf. Section 4.1) such as* $Z^\star(.)$ *is centered under* $H_{0,\, b\tilde{t}}$ *and with the mean function* $m_{\vec{t}^\star}(.)$ *of the theorem under* $H_{a\vec{t}^\star,\, b\tilde{t}}$.

The proof is given in Section 7.2. According to the proposition, our method which is based only on points of the process taken at marker positions, is not affected by epistasis. Indeed, under $H_{a\vec{t}^\star,\, b\tilde{t}}$, the mean function at marker position is the same as previously.

Figures 8 to 11 illustrate this phenomenon. The same map as previously is considered. In Figures 8 and 9, we consider two additive QTL on the chromosome : one with effect -0.6 at 10cM and the other with effect 0.8 at 70cM. To begin, in Figure 8, we consider one interaction : we have choosen to study an interaction between the two QTL. We consider two different effects for this interaction (-0.4 and 0.7). Note that the corresponding heritability is mentioned (additive+interaction). We can observe that the two additive QTL are largely

found and the number of additive QTL found is good. Then, in Figure 9, we consider this time 10 and 20 interactions (keeping the interaction between the QTL with effect -0.4). The results are still nice : the performances of our method are not affected by the interactions (as expected with the Proposition). Same conclusions hold with 4 additive QTL (see Figures 10 and 11). Note that for Figure 11, we kept the same interaction between QTL as on the left side of Figure 10, and we added other interactions.

6.4. Our method is suitable for dense map

To conclude, we would like to mention that our method is also suitable for dense map (ie a large number of genetic markers close to each other). In this case, we will perform only tests on genetic markers. In Figure 12, we consider, as previously, a chromosome of length $T = 100$cM, but genetic markers are now located every 5cM. We look for QTL every 5cM. We compare here our method and a classical LASSO method which consists of a linear model where the trait Y is the variable to explain and the regressors are the markers. In order to perform the classical LASSO, we used 0.1 as a tuning parameter instead of 5 fold cross-validation. It was a good compromise (between the QTL found and their number) since the results of the cross-validation were not good at all. According to the Figure (using the same rules to fill the table as in Section 6.1), we can see that our method gives largely better results than the classical LASSO. Note that our method is still theoretically unaffected by any interactions.

7. Proofs

7.1. Proof of the theorem

We will consider values t, $t_1^\star$, ..., $t_m^\star$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions.

Study under H_0 :
There is no QTL on the chromosome. The proof is fully given in Rabier (2010).
Nevertheless, we remind that the score test statistic for n observations verifies at position t :

$$S_n(t) = \sum_{j=1}^{n} \frac{(y_j - \mu)\ (2\ p_j(t) - 1)}{\sigma\ \sqrt{n}\ \sqrt{\mathbb{E}\left[\{2p(t) - 1\}^2\right]}} \tag{7}$$

where $\mathbb{E}\left[\{2p(t) - 1\}^2\right] = \{\alpha(t)\}^2\ +\ \{\beta(t)\}^2\ +\ 2\ \alpha(t)\ \beta(t) e^{-2(t^r - t^\ell)}$.
It will be useful for the study of the general alternative.

Study under $H_{a\vec{t}^\star}$:
There are several QTL located on the chromosome. We suppose that the QTL effects are additives and that there is no interaction between them.
In this context, the quantitative trait Y verifies :

$$Y_j = \mu\ +\ \sum_{s=1}^{m} X_j(t_s^\star)\ q_s\ +\ \sigma\varepsilon_j \tag{8}$$

where ε_j is a Gaussian white noise.
Let's introduce some notations :

- ξ : number of "Marker intervals" which contain the QTL.
  $\gamma = 1, ..., \xi$ will refer to the different intervals.
- m_γ : number of QTL in the interval γ.
  $\tau = 1, ..., m_\gamma$ refers to the τth QTL in the interval γ.
- the sth QTL on $[0, T]$, can be rewritten, $s = (\tau, \gamma) = \left\{\sum_{i=1}^{\gamma-1} m_i\right\} \ + \ \tau$

Let $\theta_{a\vec{t}^\star} = (q_1, \ ..., \ q_m, \ \mu, \ \sigma)$ and $\theta_{0\vec{t}^\star} = (0, \ ..., \ 0, \ \mu, \ \sigma)$.
After some calculations, the likelihood of $\left(Y, \ X\left\{t^{\star\ell}_{(1,1)}\right\}, \ X\left\{t^{\star r}_{(1,1)}\right\}, \ ..., \ X\left\{t^{\star\ell}_{(1,\xi)}\right\}, \ X\left\{t^{\star r}_{(1,\xi)}\right\}\right)$ with respect to the measure $\lambda \otimes N \otimes ... \otimes N$, λ being the Lebesgue measure, N the county measure on $\mathbb{N}$, verifies :

$$L^\star(\theta_{a\vec{t}^\star}) = \sum_{(u_1,...,u_m)\in\{-1,1\}^m} f_{(\mu+u_1q_1+...+u_mq_m,\sigma)}(y)$$
$$\times \left\{ \left(\prod_{\gamma=1}^{\xi} A\left\{t^{\star\ell}_{(\tau,\gamma)} \ , \ t^\star_{(\tau,\gamma)}\right\} \left[\prod_{\tau=1}^{m_\gamma-1} R\left\{t^\star_{(\tau,\gamma)} \ , \ t^\star_{(\tau+1,\gamma)}\right\}\right] A\left\{t^{\star r}_{(m_\gamma,\gamma)} \ , \ t^\star_{(m_\gamma,\gamma)}\right\} \right) g^\star(\vec{t}^\star) \right\}$$

where

$$\begin{aligned}
u_s &= u_{(\tau,\gamma)} \\
A\left\{t \ , \ t^\star_{(\tau,\gamma)}\right\} &= r\left\{t \ , \ t^\star_{(\tau,\gamma)}\right\} \ 1_{X(t)u(\tau,\gamma)=-1} \ + \ \bar{r}\left\{t \ , \ t^\star_{(\tau,\gamma)}\right\} \ 1_{X(t)u(\tau,\gamma)=1} \\
R\left\{t^\star_{(\tau,\gamma)} \ , \ t^\star_{(\tau+1,\gamma)}\right\} &= \bar{r}\left\{t^\star_{(\tau,\gamma)} \ , \ t^\star_{(\tau+1,\gamma)}\right\} \ 1_{u(\tau,\gamma)u(\tau+1,\gamma)=1} \\
&\quad + r\left\{t^\star_{(\tau,\gamma)} \ , \ t^\star_{(\tau+1,\gamma)}\right\} \ 1_{u(\tau,\gamma)u(\tau+1,\gamma)=-1} \\
g^\star(\vec{t}^\star) &= \frac{1}{2} \ \prod_{\gamma=1}^{\xi-1} D\left\{t^{\star r}_{(m_\gamma,\gamma)}, t^{\star\ell}_{(1,\gamma+1)}\right\} \\
D(t,t') &= \bar{r}(t,t') \ 1_{X(t)X(t')=1} \ + \ r(t,t') \ 1_{X(t)X(t')=-1}
\end{aligned}$$

The likelihood $L^\star_n(\theta_{a\vec{t}^\star})$ for n observations is obtained by the product of n terms as above.
Let Q_n and P_n two sequences of probability measures defined on the same space $(\Omega_n, \ \mathcal{A}_n)$. Q_n (respectively P_n) is the law corresponding to the density $L^\star_n(\theta_{a\vec{t}^\star})$ (resp $L^\star_n(\theta_{0\vec{t}^\star})$). We will call the log likelihood ratio $\log\frac{dQ_n}{dP_n}$. It verifies : $\log\frac{dQ_n}{dP_n} = \log\left\{\frac{L^\star_n(\theta_{a\vec{t}^\star})}{L^\star_n(\theta_{0\vec{t}^\star})}\right\}$.
As the model is differentiable in quadratic mean at $\theta_{a\vec{t}^\star}$ and according to the central limit theorem :

$$\log\left(\frac{dQ_n}{dP_n}\right) \overset{H_0}{\to} N(-\frac{1}{2}\vartheta^2, \ \vartheta^2) \ \text{ with } \vartheta^2 \in \mathbb{R}^{+\star}$$

By the iii) of Le Cam's first lemma, we have $Q_n \triangleleft P_n$.
Let $o_{P_{\theta_0}}(1)$ be short for a sequence of random vectors that converges to zeros in probability under H_0 (i.e. no QTL on the whole interval studied).
Besides, according to Rabier (2010) :

$$\Lambda_n(t) = \{S_n(t)\}^2 \ + \ o_{P_{\theta_0}}(1)$$

where $S_n(t)$ is given in formula (7).
Let $o_{P_{\theta_{0\vec{t}^\star}}}(1)$ be a sequence of random vectors that converges to zeros if there is no QTL at $t_1^\star$, ..., $t_m^\star$. Then, it is clear that :

$$\Lambda_n(t) = \{S_n(t)\}^2 \ + \ o_{P_{\theta_{0\vec{t}^\star}}}(1)$$

Let $o_{P_{\theta_{a\vec{t}^\star}}}(1)$ be a sequence of random vectors that converges to zeros if there are m QTL at $t_1^\star$, ..., $t_m^\star$. As $Q_n \triangleleft P_n$, according to iv) of Le Cam's first lemma :

$$\Lambda_n(t) = \{S_n(t)\}^2 \ + \ o_{P_{\theta_{a\vec{t}^\star}}}(1)$$

So, calculations can be done with the score test statistic.
According to Rabier (2010), the score test statistic at t can be obtained by a non linear interpolation :

$$S_n(t) = \frac{\alpha(t)\ S_n(t^\ell) \ + \ \beta(t)\ S_n(t^r)}{\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}$$

where $\alpha(t) = Q_t^{1,1} + Q_t^{1,-1} - 1$ and $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$.
Let $m_{\vec{t}^\star}(.)$ be the asymptotic mean function of the score process $S_n(.)$. It comes :

$$m_{\vec{t}^\star}(t) = \frac{\alpha(t)\ m_{\vec{t}^\star}(t^\ell) \ + \ \beta(t)\ m_{\vec{t}^\star}(t^r)}{\sqrt{\mathbb{E}\left[\{2p(t)-1\}^2\right]}}$$

Let calculate the quantities $m_{\vec{t}^\star}(t^\ell)$ and $m_{\vec{t}^\star}(t^r)$.
We remind that t_k refers to the location of marker k. According to Rabier (2010), the score statistic on marker k verifies :

$$S_n(t_k) = \sum_{j=1}^{n} \frac{(y_j - \mu)\ \left\{2\ 1_{X_j(t_k)=1} - 1\right\}}{\sigma\ \sqrt{n}}$$

According to formula (8) :

$$\begin{aligned} S_n(t_k) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \ \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \ + \ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \\ &= S_n^0(t_k) \ + \ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \end{aligned} \tag{9}$$

where $S_n^0(t_k)$ is the score obtained under H_0 at location t_k.
By the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} \ \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \to \mathbb{E}\left[\left\{ \sum_{s=1}^{m} X(t_s^\star)\ a_s \right\} \left\{2\ 1_{X(t_k)=1} - 1\right\} \right]$$

According to Rabier (2010), we have :

$$\mathbb{E}\left[\left\{\sum_{s=1}^{m} X(t_s^\star)\ a_s\right\}\left\{2\ 1_{X(t_k)=1}-1\right\}\right]=\sum_{s=1}^{m} a_s\ e^{-2|t_s^\star-t_k|}$$

It comes :

$$m_{\vec{t}^\star}(t_k)=\frac{1}{\sigma}\sum_{s=1}^{m} a_s\ e^{-2|t_s^\star-t_k|}$$

As a consequence :

$$m_{\vec{t}^\star}(t^\ell)=\frac{1}{\sigma}\sum_{s=1}^{m} a_s\ e^{-2|t_s^\star-t^\ell|}\quad,\quad m_{\vec{t}^\star}(t^r)=\frac{1}{\sigma}\sum_{s=1}^{m} a_s\ e^{-2|t_s^\star-t^r|}$$

Weak convergence of the score process :
The proof is exactly the same as in Rabier (2010).

7.2. Proof of the proposition

$\tilde{m}$ is the number of interactions and $\tilde{q}_s$ the effect of the sth interaction. The loci corresponding to the sth interaction are $\tilde{t}_{2s}$ and $\tilde{t}_{2s-1}$. In this context, the quantitative trait Y verifies :

$$Y=\mu\ +\ \sum_{s=1}^{m} X(t_s^\star)\ q_s\ +\ \sum_{s=1}^{\tilde{m}} X(\tilde{t}_{2s-1})X(\tilde{t}_{2s})\ \tilde{q}_s\ +\ \sigma\varepsilon \tag{10}$$

where ε is a Gaussian white noise.
We will consider values of $\tilde{t}_1$, ..., $\tilde{t}_{2\tilde{m}}$ and $t_1^\star$, ..., $t_m^\star$ distinct of marker positions, and the result will be prolonged by continuity.
Let $o_{P_{\theta_{0\vec{t}^\star,0\tilde{t}}}}(1)$ be a sequence of random vectors that converges to zeros if there is no additive QTL at $t_1^\star$, ..., $t_m^\star$ and no interactions between loci $\tilde{t}_1$ and $\tilde{t}_2$, ..., no interactions between loci $\tilde{t}_{2\tilde{m}-1}$ and $\tilde{t}_{2\tilde{m}}$. In the same way as in the proof of the theorem, it is clear that :

$$\Lambda_n(t_k)=\{S_n(t_k)\}^2\ +\ o_{P_{\theta_{0\vec{t}^\star,0\tilde{t}}}}(1)$$

where $S_n(t_k)$ is given in formula (5) of Section 5.2.
In order to adapt the proof of the theorem, we just have to consider the likelihood of Y and the flanking markers of the additive QTL (as previously) but we have to add the flanking markers of $\tilde{t}_1$, ...,$\tilde{t}_{2\tilde{m}}$. The model is still differentiable in quadratic mean.
Let $o_{P_{\theta_{a\vec{t}^\star,b\tilde{t}}}}(1)$ be a sequence of random vectors that converges to zeros if there are m additive QTL at $t_1^\star$, ..., $t_m^\star$ and $\tilde{m}$ interactions : loci $\tilde{t}_1$ and $\tilde{t}_2$, ..., loci $\tilde{t}_{2\tilde{m}-1}$ and $\tilde{t}_{2\tilde{m}}$. Then, according to iv) of Le Cam's first lemma :

$$\Lambda_n(t_k)=\{S_n(t_k)\}^2\ +\ o_{P_{\theta_{a\vec{t}^\star,b\tilde{t}}}}(1)$$

According to formula (10), we have :

$$\begin{aligned} S_n(t_k) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \ \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \ + \ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \\ &+ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{\tilde{m}} X_j(\tilde{t}_{2s-1}) X_j(\tilde{t}_{2s})\ b_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \\ &= S_n^0(t_k) \ + \ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{m} X_j(t_s^\star)\ a_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \\ &+ \frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{\tilde{m}} X_j(\tilde{t}_{2s-1}) X_j(\tilde{t}_{2s})\ b_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \end{aligned} \tag{11}$$

where $S_n^0(t_k)$ is the score obtained under the null hypothesis that there is no additive QTL and no interactions on $[0, T]$ (same S_n^0 as in formula (9) of the proof of the theorem). According to the proof of the theorem, we have $\frac{1}{\sigma n} \sum_{j=1}^{n} \{\sum_{s=1}^{m} X_j(t_s^\star)\ a_s\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\}$ which tends to $m_{\vec{t}^\star}(t_k)$. Besides,

$$\frac{1}{\sigma n} \sum_{j=1}^{n} \left\{ \sum_{s=1}^{\tilde{m}} X_j(\tilde{t}_{2s-1}) X_j(\tilde{t}_{2s})\ b_s \right\} \left\{2\ 1_{X_j(t_k)=1} - 1\right\} \to \mathbb{E}\left[\left\{ \sum_{s=1}^{\tilde{m}} X(\tilde{t}_{2s-1}) X_j(\tilde{t}_{2s}) b_s \right\} \left\{2\ 1_{X(t_k)=1} - 1\right\} \right]$$

We have :

$$\mathbb{E}\left[X(\tilde{t}_{2s-1}) X(\tilde{t}_{2s}) \left\{2\ 1_{X(t_k)=1} - 1\right\}\right] = 2\ \mathbb{E}\left[X(\tilde{t}_{2s-1}) X(\tilde{t}_{2s}) 1_{X(t_k)=1}\right] - e^{-2|\tilde{t}_{2s} - \tilde{t}_{2s-1}|}$$

If $t_k < \tilde{t}_{2s-1} < \tilde{t}_{2s}$, then :

$$\mathbb{E}\left[X(\tilde{t}_{2s-1}) X(\tilde{t}_{2s}) \left\{2\ 1_{X(t_k)=1} - 1\right\}\right] = 0$$

If $\tilde{t}_{2s-1} < t_k < \tilde{t}_{2s}$, then :

$$\mathbb{E}\left[X(\tilde{t}_{2s-1}) X(\tilde{t}_{2s}) \left\{2\ 1_{X(t_k)=1} - 1\right\}\right] = 0$$

As a consequence :

$$\mathbb{E}\left[X(\tilde{t}_{2s-1}) X(\tilde{t}_{2s}) \left\{2\ 1_{X(t_k)=1} - 1\right\}\right] = 0$$

It concludes the proof for under $H_{a\vec{t}^\star,\ b\tilde{t}}$. In order to obtain the result under $H_{0,\ b\tilde{t}}$, we just have to deal with contiguity, considering the likelihood of Y and only the flanking markers of $\tilde{t}_1$, ...,$\tilde{t}_{2\tilde{m}}$ (ie the loci for the interactions). Then, we do the same calculations as in formula (11) but this time there is not anymore the additive term (ie the second term). It concludes the proof of the Proposition.

8. Acknowledgements

The authors thank Jean-Michel Elsen for having proposed this subject of research and fruitful discussions. This work has been supported by the Animal Genetic Department of the French National Institute for Agricultural Research, SABRE, and the National Center for Scientific Research.

locations (in cM)	(10 ; 70)			(30 ; 80)		
QTL effects	$(-0.6\ ;\ 0.8)$	$(-0.8\ ;\ 0.8)$	$(0.4\ ;\ -0.6)$	$(0.6\ ;\ 0.6)$	$(0.6\ ;\ 0.8)$	$(0.6\ ;\ 0.4)$
h^2	42%	47%	27%	50%	57%	41%
QTL found	(88% ; 100%)	(100% ; 94%)	(75% ; 96%)	(97% ; 98%)	(96% ; 100%)	(100% ; 94%)
nb of QTL found	2.49	2.71	2.46	2.49	2.42	2.68

Fig. 4. Percentage of QTL truly identified (QTL found) and number of QTL found (nb of QTL found) as a function of the QTL effects, their locations (h^2 refers to the heritablities). 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). 2 QTL lie on the chromosome. We look for QTL every 5cM. In the notation (a, b), a refers to the first QTL and b to the second one.

nb of QTL	3		4	
locations (in cM)	(10 ; 40 ; 90)		(10 ; 50 ; 70 ; 90)	
QTL effects	$(-0.6\ ;\ -0.6\ ;\ 0.4)$	$(-0.6\ ;\ -0.6\ ;\ 0.6)$	$(0.4\ ;\ 0.4\ ;\ 0.4\ ;\ 0.4)$	$(0.6\ ;\ 0.6\ ;\ 0.6\ ;\ 0.6)$
h^2	50%	52%	61%	78%
QTL found	(94% ; 85% ; 56%)	(94% ; 86% ; 86%)	(77% ; 71% ; 96% ; 81%)	(83% ; 66% ; 97% ; 81%)
nb of QTL found	3.54	3.70	4.21	4.24

Fig. 5. Percentage of QTL truly identified (QTL found) and number of QTL found (nb of QTL found) as a function of the number of QTL, their effects and their locations. 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). We look for QTL every 5cM.

	nb of QTL	2	2	3
	locations (in cM)	(10 ; 70)	(10 ; 70)	(10 ; 40 ; 80)
	QTL effects	(-0.6 ; 0.8)	(0.4 ; -0.6)	(0.4 ; 0.7 ; -0.8)
	h^2	42%	27%	48%
this paper	QTL found	(88% ; 100%)	(75% ; 96%)	(67% ; 87% ; 100%)
	nb of QTL found	2.49	2.46	3.53
$CIM_{Shuff}(20)$	QTL found	(98% ; 28%)	(81% ; 95%)	(79% ; 79% ; 71%)
	nb of QTL found	4.36	3.26	4.92
$CIM_{H_0}(20)$	QTL found	(73% ; 97%)	(9% ; 57%)	(14% ; 70% ; 56%)
	nb of QTL found	1.84	0.7	3.99
$CIM_{Shuff}(40)$	QTL found	(89% ; 87%)	(76% ; 71%)	(74% ; 100% ; 100%)
	nb of QTL found	4.86	4.38	4.97
$CIM_{H_0}(40)$	QTL found	(69% ; 77%)	(13% ; 48%)	(6% ; 100% ; 98%)
	nb of QTL found	3.29	1.70	4.08

Fig. 6. Percentage of QTL truly identified (QTL found) and number of QTL found (nb of QTL found) as a function of the number of QTL, their effects, their locations and the method. CIM_{Shuff} (resp. CIM_{H_0}) refers to CIM using a permutation threshold (resp. threshold obtained with no QTL). $CIM(20)$ (resp. $CIM(40)$) refers to CIM with markers considered as covariates if they do not belong to a window size of 20cM (resp. 40cM) of the position tested. 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). We look for QTL every 5cM.

	QTL effects	(-0.4 ; -0.7 ; 0.9 ; 0.8)	(-0.8 ; -0.8 ; 0.8 ; 0.8)	(-0.4 ; -0.4 ; 0.6 ; 0.8)
	h^2	66%	70%	58%
this paper	QTL found	(72% ; 68% ; 77% ; 100%)	(97% ; 83% ; 57% ; 100%)	(78% ; 54% ; 57% ; 100%)
	nb of QTL found	4.08	3.94	3.55
$CIM_{Shuff}(20)$	QTL found	(59% ; 93% ; 96% ; 98%)	(90% ; 96% ; 75% ; 96%)	(53% ; 56% ; 86% ; 98%)
	nb of QTL found	4.87	5.00	4.52
$CIM_{H_0}(20)$	QTL found	(02% ; 71% ; 95% ; 97%)	(95% ; 77% ; 86% ; 93%)	(09% ; 06% ; 75% ; 100%)
	nb of QTL found	3.71	4.82	2.54
$CIM_{Shuff}(40)$	QTL found	(63% ; 100% ; 59% ; 00%)	(91% ; 100% ; 48% ; 24%)	(68% ; 89% ; 41% ; 18%)
	nb of QTL found	4.81	5.00	4.82
$CIM_{H_0}(40)$	QTL found	(03% ; 84% ; 58% ; 00%)	(86% ; 98% ; 52% ; 30%)	(11% ; 32% ; 46% ; 14%)
	nb of QTL found	3.79	4.94	3.20

Fig. 7. Percentage of QTL truly identified (QTL found) and number of QTL found (nb of QTL found) as a function of the QTL effects and the method. CIM_{Shuff} (resp. CIM_{H_0}) refers to CIM using a permutation threshold (resp. threshold obtained with no QTL). $CIM(20)$ (resp. $CIM(40)$) refers to CIM with markers considered as covariates if they do not belong to a window size of 20cM (resp. 40cM) of the position tested. 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). 4 QTL lie on the chromosome at 10cM, 40cM, 70cM and 90cM. We look for QTL every 5cM.

effect of the interaction between the two QTL	-0.4	0.7
h^2	47%	54%
additive QTL found	$(86\% \ ; \ 98\%)$	$(80\% \ ; \ 93\%)$
nb of additive QTL found	2.61	2.53

Fig. 8. Percentage of additive QTL truly identified (additive QTL found) and number of additive QTL found (nb of additive QTL found) as a function of the effect of the interaction. 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). 2 additive QTL lie on the chromosome with effects -0.6 at 10cM and 0.8 at 70cM. We look for additive QTL every 5cM.

nb of interactions	10	20
h^2	54%	59%
additive QTL found	$(82\% \ ; \ 93\%)$	$(74\% \ ; \ 91\%)$
nb of additive QTL found	2.60	2.57

Fig. 9. Percentage of additive QTL truly identified (additive QTL found) and number of additive QTL found (nb of additive QTL found) as a function of the number of interactions and as a function of the heritability (h^2) considered. 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). 2 additive QTL lie on the chromosome with effects -0.6 at 10cM and 0.8 at 70cM. We look for additive QTL every 5cM.

interactions between QTL	$(1$ and $3 \ ; \ 2$ and $4)$	$(1$ and $4 \ ; \ 2$ and $3)$
effects of the interactions	$(-0.4 \ ; \ -0.6)$	$(-0.4 \ ; \ -0.6)$
h^2	71%	75%
additive QTL found	$(61\% \ ; \ 76\% \ ; \ 64\% \ ; \ 100\%)$	$(66\% \ ; \ 70\% \ ; \ 65\% \ ; \ 100\%)$
nb of additive QTL found	3.79	3.86

Fig. 10. Percentage of additive QTL truly identified (additive QTL found) and number of additive QTL found (nb of additive QTL found) as a function of the interactions considered and their effects. 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). 4 additive QTL lie on the chromosome with effects -0.4 at 10cM, -0.7 at 40cM, 0.9 at 70cM, 0.8 at 90cM. We look for additive QTL every 5cM.

number of interactions	6	10
h^2	75%	77%
additive QTL found	(72% ; 79% ; 61% ; 100%)	(58% ; 65% ; 57% ; 100%)
nb of additive QTL	3.86	3.67

Fig. 11. Percentage of additive QTL truly identified (additive QTL found) and number of additive QTL found (nb of additive QTL found) as a function of the number of interactions and as a function of the heritability (h^2). 100 populations of $n = 320$ individuals are considered. 6 genetic markers are equally spaced every 20cM ($T = 100$cM). 4 additive QTL lie on the chromosome with effects -0.4 at 10cM, -0.7 at 40cM, 0.9 at 70cM, 0.8 at 90cM. We look for additive QTL every 5cM.

	nb of interactions	0	10	20
	h^2	48%	60%	64%
this paper	additive QTL found	(100% ; 88% ; 100%)	(100% ; 76% ; 93%)	(99% ; 71% ; 91%)
	nb of additive QTL found	3.44	3.13	3.05
LASSO	additive QTL found	(83% ; 67% ; 72%)	(82% ; 73% ; 71%)	(88% ; 70% ; 71%)
	nb of additive QTL found	5.67	5.95	5.76

Fig. 12. Percentage of additive QTL truly identified (additive QTL found) and number of additive QTL found (nb of additive QTL found) as a function of the number of interactions, the heritability (h^2) and the method considered. 100 populations of $n = 320$ individuals are considered. 21 genetic markers are equally spaced every 5cM ($T = 100$cM). 3 additive QTL lie on the chromosome with effects -0.8 at 5cM, 0.8 at 45cM, -0.8 at 70cM. We look for additive QTL every 5cM.

References

Broman, K. W., Wu, H., Sen, S., Churchill, G. A., (2003) R/qtl : QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889-890.

Broman, K.W., Speed, T., (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society Serie B*, **64**, 641-656.

Chang, M. N., Wu, R., Wu, S. S., Casella, G., (2009) Score statistics for mapping quantitative trait loci. *Statistical Application in Genetics and Molecular Biology*, **8(1)**, 16.

Churchill, G.A., Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping *Genetics*, **138**, 963-971.

Cierco, C., (1998) Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*, **31**, 261-285.

Feingold, E., Brown, P.O., Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of idendity by descent. *Am. J. Human. Genet.*, **53**, 234-251.

Haldane, J.B.S (1919) The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*, **8**, 299-309.

Hastie, T., Tibshirani R., Friedman J. (2001) *The elements of statitical learning*. Springer.

Jansen, R.C. (1993) Interval Mapping of multiple Quantitative Trait Loci. *Genetics*, **135**, 205-211.

Lander, E.S., Botstein, D., (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138**, 235-240.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*, Springer.

Martinez, O., Curnow, R.N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers Theoretical and Applied Genetics **85**, 480-488.

Piepho, H-P. (2001). A quick method for computing approximate thresholds fo quantitative trait loci detection Genetics **157**, 425-432.

Rabier, C-E. (2010) *PhD thesis*, Université Toulouse 3, Paul Sabatier.

Rebaï, A., Goffinet, B., Mangin, B. (1994) Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235-240.

Rebaï, A., Goffinet, B., Mangin, B. (1995) Comparing power of different methods for QTL detection. *Biometrics*, **51**, 87-99.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society - B*, **58**, **1**, 267-288.

Van der Vaart, A.W. (1998) *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

Wu, R., MA, C.X., Casella, G. (2007) *Statistical Genetics of Quantitative Traits*, Springer

Zeng, Z-B. (1994) Precision Mapping of Quantitative Trait Loci, *Genetics*, **136**, 1457-1468.